

# Big Data and Hadoop essentials

<https://www.udemy.com/big-data-and-hadoop-essentials-free-tutorial/?dtcode=D8F9u2F1jMLu>

## Section 1, Lecture 2

The screenshot shows a video player interface. The slide content is as follows:

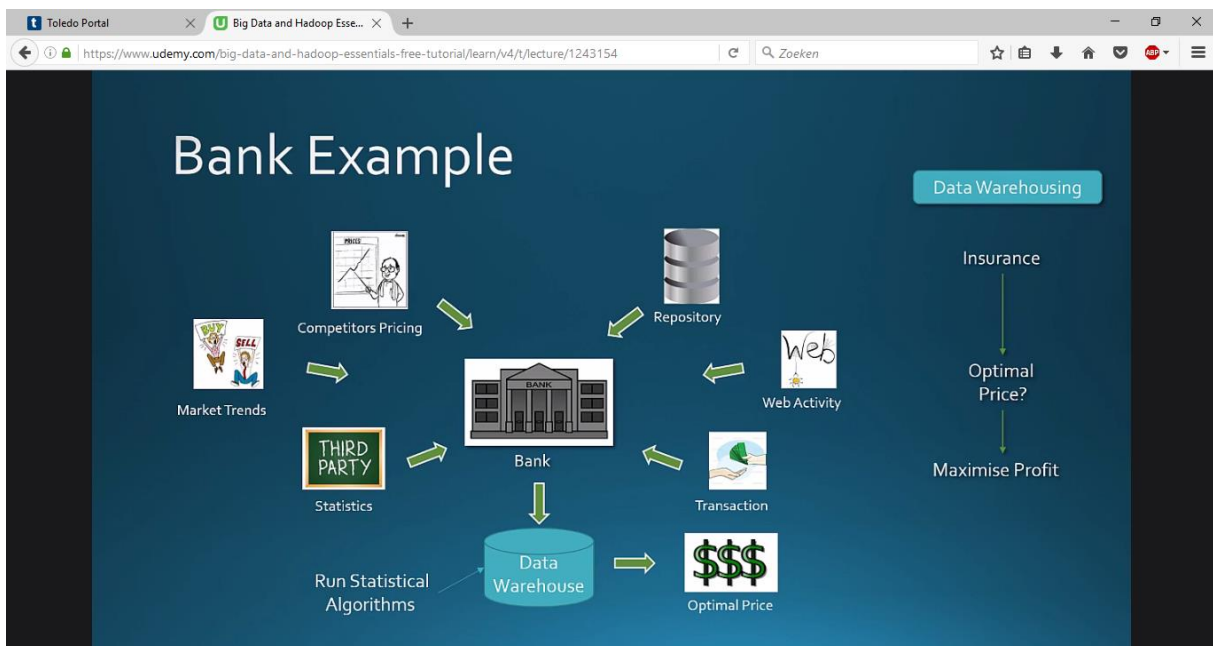
Lesson 1 Understanding Big Data  
Section 1, Lecture 2

### The hype around Big Data

- ▶ Facebook, Twitter, Google generating petabytes of data everyday.
- ▶ Hadron Collider project discarding large amount of data as they won't be able to analyse. Hoping that they haven't thrown anything valuable.

Interesting facts but .... Why is Big Data important?

1:27 / 7:57



Big Data – Text book definition

*"Big data are a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications"*

-White Tom, Definitive Guide

Volume      Variety      Velocity

### Section 1, Lecture 3

Lesson 2 Future of Data  
Section 1, Lecture 3

Go to Dashboard

## Example – Digital Nervous System

The diagram illustrates a 'Digital Nervous System' where various data sources feed into a central 'Bank' icon. A green arrow points from the 'Bank' to a box containing '\$\$\$', which is labeled 'Optimal Price'. The data sources include:

- Competitors Pricing (with a line graph icon)
- Market Trends (with a newspaper icon)
- Bank (with a building icon)
- Web Activity (with a 'Web' icon)
- Transaction (with a hand holding a coin icon)
- Mobile Alert with Travel insurance (with an illustration of a person at a laptop)
- Statistics (with a 'THIRD PARTY' sign icon)
- Repository (with a database cylinder icon)

Optimal Price  
Browse Q&A    Add Bookmark    Continue >

2:21/5:32

## Section 1, Lecture 4



Hadoop Distributed File System and Hadoop MapReduce were inspired from Google File System and Google MapReduce Papers.

✓ True

Good job, please continue to the next question.

False

Hadoop spawned of Nutch with an idea that distributed framework architecture could be used for more purposes than just to solve search engine algorithms.

✓ True

Good job, please continue to the next question.

False

Apache software foundation has released the fundamental hadoop open source versions.

✓ True  
Good job, please continue to the next question.

False

Great job! You are ready to move on to the next lecture.  
You got 3 out of 3 correct on the first attempt.

- ✓ What you know ⓘ
  - Hadoop Distributed File System and Hadoop MapReduce were inspired from Google...
  - Hadoop spawned of Nutch with an idea that distributed framework architecture coul...
  - Apache software foundation has released the fundamental hadoop open source ver...

Section 2, Lecture 5

Ox and the load

0:51 / 7:44

Continue >

Browse Q&A Add Bookmark

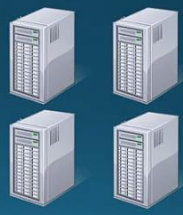

Big Data and Hadoop Esse... x +

https://www.udemy.com/big-data-and-hadoop-essentials-free-tutorial/learn/v4/t/lecture/1243162 Zoeken

Lesson 4 - Hadoop Magic  
Section 2, Lecture 5

# Distributed Computing

Go to Dashboard



Price Advantage:

1. Clusters use commodity hardware, cheaper than one expensive server.
2. Software License is free.

1:34 / 7:44

15 1x 15

Browse Q&A Add Bookmark Continue >

Big Data and Hadoop ... x +

https://www.udemy.com/big-data-and-hadoop-essentials-free-tutorial/learn/v4/t/lecture/1243162 Zoeken

# The new Fundamentals

- Moving the code to data
- Use of Commodity Hardware and Open Source Software against expensive proprietary software on expensive custom Hardware.
- On read schema.

A pseudo organisation is looking to create a data warehouse system to run data analytics to get deeper insights of the customer behaviour. Which one would be more recommended between the two implementations :

**Implementation 1:**

Hardware Server (costing around 50K), Proprietary BI softwares like Cognos and Oracle for RDBMS etc

**Implementation 2:**

Cluster of Computers (costing 1K each), Open source softwares like Hadoop, Mahout and NoSql Database etc,

1 Implementation 1 is the best suited as Hadoop technologies is new and is not much efficient.

2 Implementation 2 is better suited as Hadoop is the new way to go.

3 It depends on the situation. Implementation 2 suites better for the batch processing requirements and for data sizes in PB and beyond.

4 On the other hand implementation 1, would suite better if the data to be processed would be in the range of a few GBs and quick real time environment is needed.

Which one of the following is ***not*** a new fundamental idea brought in by hadoop

Move the code to data rather than the data to code.

Use of commodity hardware, rather than expensive custom hardware.

To have schema on read.

✓ By looking at the data around, a business can have deeper insights to customer.  
**Good job, please continue to the next question.**

Add Answer

Big Data and Hadoop Esse... x +

https://www.udemy.com/big-data-and-hadoop-essentials-free-tutorial/learn/v4/t/quiz/59054

Go to Dashboard

Great job! You are ready to move on to the next lecture.

You got 2 out of 2 correct on the first attempt.

✓ What you know ⓘ

A pseudo organisation is looking to create a data warehouse system to run data anal...

Which one of the following is not a new fundamental idea brought in by hadoop



## Section 2, Lecture 6

Big Data and Hadoop ...

https://www.udemy.com/big-data-and-hadoop-essentials-free-tutorial/learn/v4/t/lecture/1243164

Lesson 5 - Hadoop Ecosystem  
Section 2, Lecture 6

# Hadoop Ecosystem

Go to Dashboard

The diagram illustrates the Hadoop ecosystem. At the top, 'Yahoo' and 'Facebook' are connected to 'Pig' and 'Hive' respectively. Below them are 'MapReduce' and 'HBase'. At the bottom is 'HDFS'. To the right, 'Sqoop/Flume' is connected to 'HDFS' and 'Structured Stores'.

2:37/4:07

Continue >

Big Data and Hadoop ...

https://www.udemy.com/big-data-and-hadoop-essentials-free-tutorial/learn/v4/t/lecture/1243164

Lesson 5 - Hadoop Ecosystem  
Section 2, Lecture 6

# Hadoop Ecosystem

Go to Dashboard

The diagram illustrates the Hadoop ecosystem. At the top, 'Oozie' is connected to 'Pig' and 'Hive'. Below them are 'MapReduce' and 'HBase'. At the bottom is 'HDFS'. To the right, 'Sqoop/Flume' is connected to 'HDFS' and 'Structured Stores'. Below 'Sqoop/Flume' are 'Storm', 'Chukwa', and 'Kafka', which are connected to 'HDFS'. 'Storm' is connected to 'Structured Stores', 'Chukwa' to 'Log collection', and 'Kafka' to 'Message broker'.

3:13/4:07

Continue >



Hadoop ecosystem projects were independently developed and hence have a lot of compatibility issue and thus manual installation from scratch is not recommended. Open source packages provided by vendors is more recommended.

✓ True  
Good job, please continue to the next question.

False

### Section 2, Lecture 7

The screenshot shows a video player interface for a UDEMY course. The video content is a slide titled "Simpler Vs Complex Algorithms". The slide features two scatter plots side-by-side. The left plot is labeled "Complex Algorithm on a small dataset" and shows a small number of data points with a red regression line. The right plot is labeled "Simple Algorithm on a large dataset" and shows a larger number of data points with a red regression line. To the right of the plots, there is a list of two points:

1. Complex Algorithms needs to be correctly sensitive to weak correlations.
2. Complex Algorithms are thus difficult to code and design.

The video player interface includes a progress bar at the bottom, a "Continue" button, and a timestamp of 2:08/8:39.

Big Data and Hadoop ... x +

https://www.udemy.com/big-data-and-hadoop-essentials-free-tutorial/learn/v4/t/lecture/1243166

Zoeken

# Data Engineer Vs Data Scientist

	Data Engineer	Data Scientist
Role	To engineer software solutions.	To solve business problems using data.
Skills	More of programing and technical skills and ability to architect technical solutions.	Strong of Mathematical Skills and understanding of statistical Models.

Big Data and Hadoop ... x +

https://www.udemy.com/big-data-and-hadoop-essentials-free-tutorial/learn/v4/t/lecture/1243166

Zoeken

# Hadoop Vendors

Vendor	Notes
Apache	-> Skeleton Version -> All the ecosystems need to be additionally installed.
Cloudera	-> Important ecosystem members included. -> Few Proprietary tools like Enterprise Manager.
MAPR Technologies	-> Proprietary Hadoop code written in C. -> Integrated with Hadoop ecosystem members.
Hortonworks	-> Based out of Apache hadoop. -> Supports .NET framework
EMC <sup>2</sup>	-> Launches Hadoop Distribution: Pivotal HD

Big Data and Hadoop Esse... x +

https://www.udemy.com/big-data-and-hadoop-essentials-free-tutorial/learn/v4/t/quiz/59066

Go to Dashboard

Great job! You are ready to move on to the next lecture.

You got 2 out of 2 correct on the first attempt.

✔ What you know ⓘ

- Simple algorithms on a large dataset produces much accurate results than a comple...
- If there is an organisation which is based out and have core software on .NET frame...

Section 2, Lecture 8

